Editors' note: *This series addresses topics that affect epidemiologists across a range of specialties. Commentaries are first invited as talks at symposia organized by the Editors. This article was originally presented at the 2006 Congress of Epidemiology in Seattle.*

The Evolution of Epidemiologic Research

From Cottage Industry to "Big" Science

Robert N. Hoover

The term "big science" was first coined to describe several changes in the way science (particularly physics) was conducted in industrial nations during and after World War II. These changes included the assembly of big staffs (usually multidisciplinary) to work in big laboratories, using big machines, all requiring big budgets. The proximal stimulus for this was the belief that this was the only way to make rapid progress in developing a variety of critical defense-related products for the war effort, including the proximity fuse, radar, and the atomic bomb. This general approach, however, continued after the war, extending to particle research in general, and into other areas of physics as well (eg, laser technology). As time went on, other disciplines developed significant "big" science components as well, perhaps most notably astronomy. These trends have only recently entered the biologic sciences with the most graphic example being the changes in genomic research over the last 15 years.

So how did "big" science come to epidemiology? The process has been somewhat different for classic epidemiology and molecular epidemiology. For the classic approaches, the need developed in a gradual and subtle manner. Up until 25 to 30 years ago, most epidemiologic research could be described as a cottage industry. There were a number of programs, but most were comprised of a small number of investigators with expertise largely limited to epidemiology, medicine, and biostatistics. Each investigator worked on his own studies, which were generally small (a study of 300 cases of a disease was viewed as a large study) using a limited study team and run on a relatively small budget. The principal investigator tended to design a study, develop the interview and abstract forms, train and supervise the data collectors (often doing much of it himself), code the data, and conduct the analyses on a calculator or perhaps through simple computer programs he could write himself.

Although all this sounds rather quaint to young investigators today, it actually worked quite well, and this was a productive era for meaningful epidemiologic discoveries. The success of these approaches was largely due to the objectives of the investigations. For the most part, the interest was in identifying large risks associated with obvious and easily measured exposures with a focus only on main effects. In my own area of cancer epidemiology, this was the era when many of the major risks with tobacco and alcohol consumption, ionizing radiation, occupational chemical exposures, medicinal agents, and reproductive factors were elucidated. Although many studies continue to focus on these objectives, over the last 30 years, there has been an increasing interest in identifying relatively low levels of risk, often associated with low-level or difficult-to-measure exposures, and a concern about effects in subgroups (effect modification). This has resulted in the need for larger studies (now a study of 300 cases is often considered small), larger, more interdisciplinary study teams, and sophisticated, technology-intensive analytic methods—all of which require substantially larger budgets.

From the Division of Cancer Epidemiology and Genetics, National Cancer Institute, Department of Health and Human Services, Bethesda, MD.
Editors' note: *Related papers appear on pages 1, 9, and 18.*
Correspondence: Robert N. Hoover, Director, Epidemiology and Biostatistics Program, Division of Cancer Epidemiology and Genetics, National Cancer Institute, 6120 Executive Blvd., EPS 8094, Bethesda, MD 20852; E-mail: hooverr@mail.nih.gov.

ISSN: 1044-3983/07/1801-0013
DOI: 10.1097/01.ede.0000249532.81073.b2

As noted, the trend has been gradual but has been going on for some time. I first encountered this in my own research in the late 1970s. At that time, laboratory animal research had raised the possibility that the artificial sweetener saccharin might be a urinary bladder carcinogen. One epidemiologic study[1] had also found a significant risk (odds ratio = 1.6) but limited to males. Because of the widespread exposure, there was great concern and urgency for more reliable data. In response, we were able to develop a collaborative study[2] involving 10 population-based cancer registries funded by the National Cancer Institute. In 1 year's time, we were able to interview 3,000 subjects with bladder cancer and 5,800 population-based controls. Subsequent analysis eased concerns about any consequential risk, either overall or limited to one gender (Table 1). As time has gone on, these types of large investigations for some disease–exposure concerns have become more the norm rather than the exception.

If the need for, and development of, "big" science in classic epidemiology has been subtle and gradual, it has been anything but this for molecular epidemiology. The advent of major advances in molecular science and technology has offered the epidemiologist a variety of remarkable opportunities to overcome some of the weaknesses of more classic approaches, particularly in the areas of measurement of exposures and outcomes, detection of susceptible subgroups, and insights into the biologic mechanisms of disease. Each of these opportunities impacts on issues of "big" science, but for purposes of illustration, I focus only on assessment of susceptibility, specifically genetic susceptibility.

For some time, there has been avid interest in identifying susceptibility genes for various diseases.[3] This refers specifically to polymorphisms in genes associated with a relatively low penetrance of disease, but which are thought to function primarily by modifying the effects of disease-causing exposures. Incredible advances in genetic science and technology over the last 15 years have now given us the opportunity to make comprehensive assessments of such genetic risk factors an integral part of our epidemiologic investigations. The enthusiasm to do so stems not only from the opportunity to identify susceptibility risk factors, but also to explore gene–environment interactions. This gives us the ability to identify high- and low-risk populations, to gain insights into the actual mechanisms of pathogenesis, and to identify previously unrecognized pathogens.

Although the opportunities are great, and thus the enthusiasm for pursuing such studies is very high, there is a dark side to these opportunities. Included are issues of study size, the opportunity for chance findings, and enhanced opportunities for bias. The study size and chance issues are the most germane to a discussion of "big" science. If your interest is in assessment of interactions, it is almost a given that you need to conduct a large study (Fig. 1).[4] To have 80% power to detect even a strong (2-fold) interaction, between an exposure experienced by 10% to 20% of the population and a genotype present in 10% of the population, you would need over 1,500 cases and controls. This is if you are measuring the exposure perfectly. If you are measuring the exposure with the kinds of sensitivity common in epidemiology, the sample size requirements rapidly escalate into the several thousands.

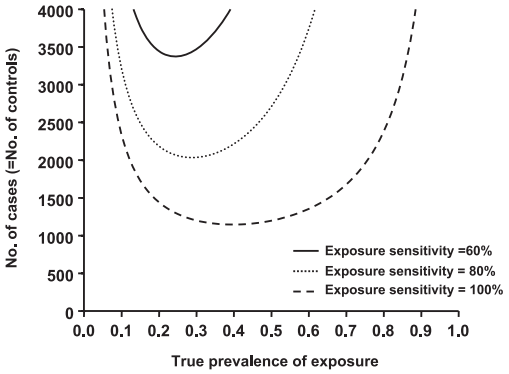

FIGURE 1. Sample size requirements for detection of a 2-fold interaction in disease risk between a genotype of 10% prevalence and an exposure by prevalence of exposure and sensitivity of its measurement (power = 0.80, $\alpha = 0.05$, specificity of measurement of exposure = 100%).[4]

With respect to the role of chance, when dealing with genetic variants, issues surrounding multiple comparisons and the possibilities for false-positive findings reach a level almost beyond comprehension. We now know there are approximately 24,000 genes in the human genome with a current estimate of around 8 million possible common variants (single nucleotide polymorphisms [SNPs]). When one considers that interest is not only in the effect of a single SNP, but in variations in pathways (multiple genes considered together), and particularly in the interactions of genes or pathways with multiple environmental and lifestyle exposures, the possible comparisons become legion, and the probability that any finding is in reality a false-positive becomes extraordinarily high. Handling this challenge then translates into not only one large study, but multiple such ones to provide opportunities for reliable replication.

This is an interesting theory, but are these challenges actual barriers in practice? One only has to review a small sample of the relevant literature from the last several years to see that these are enormous impediments to progress. Many have rushed to apply the new genomic possibilities in a variety of contexts. The result has been a blizzard of positive findings for individual genes and gene–gene and gene–environment interactions, most coming from under-

TABLE 1. Odds Ratios for Ever-Use of Artificial Sweeteners and Bladder Cancer Risk[2]

	OR	(95% CI)
Total	1.01	(0.92–1.11)
Men	0.99	(0.89–1.10)
Women	1.07	(0.89–1.29)

powered studies, findings that have been found to be largely nonreproducible. Perhaps instructive is the example of cigarette smoking, genetic susceptibility, and breast cancer risk. Since 1995, 50 studies have searched for such relationships for a total of 11 genes. A recent review and meta-analysis[5] found some evidence of consistency for a couple of genetic variants but concluded, "however, interpretation of the available literature is complicated by methodologic limitations, including small sample size, . . . which likely contributed to the inconsistent findings. These methodologic issues should be addressed in future studies . . ." If looking over a 10-year period in 50 studies for a gene–environment interaction between a common exposure and common candidate genetic variants for risk of a common cancer leads to no firm conclusions, we clearly need to think about a more coherent approach.

This example, as noted, deals with the candidate gene approach—genes chosen a priori to be investigated because of prior suspicion of their involvement. How much more problematic will be our attempts to identify and pursue genetic susceptibility based on the agnostic approaches involved in whole genome analysis? The convergence of knowledge of the human genome from the Human Genome Project, knowledge of how many of the variants (SNPs) are highly correlated with each other from the HapMap Project, the development of Dense SNP Detection Technologies, and the presence of large case–control and cohort studies with DNA collection has made it possible to consider surveying all 8 million SNPs, or markers for them, to identify those associated with a specific disease. Many of these whole genome analysis studies will be done over the next several years, assaying for 500,000 or more SNPs as markers across the entire genome of each case and control. The tradeoff between adequate power to ensure that all true positives are identified, with the accompanying production of false-positives, is truly daunting.[6] In a whole genome analysis of 500,000 SNPs of 1,200 cases and 1,200 controls, to achieve reasonable power to identify a true-positive SNP of 10% prevalence associated with an odds ratio of 1.4, one would need to choose a *P* value that would also retain 20,000+ false-positives. If it is thought that very few SNPs will be causally related to any specific disease risk (many predict 10–30 as a maximum), then you are confronted with how to identify these 10 to 30 of the 20,000+. It should also be recognized that this is just to identify a main effect on disease risk. If environmental variables are introduced into this agnostic genetic approach, and gene–environment interactions are pursued at this discovery level, the prospects become truly draconian.

Clearly, some of the marvelous opportunities brought to epidemiology with the incorporation of molecular science come with enormous challenges that many believe require the development of new research paradigms, and many of these new paradigms incorporate major elements of "big science." Indeed, meeting these challenges has been the subject of much discussion. It has been generally agreed that very large studies are needed; that they need to be rigorously designed, conducted, and analyzed; that validation will be required in more than one study; and that ideally this should occur in diverse groups. It has also be recognized that, because of the size and resources required, only a limited number of these efforts can be supported, so the ones that move forward will have to be coordinated with, and provide access to, the larger epidemiologic and biomedical communities. Several alternative ways of achieving this program have been proposed, with the 2 most prominent being the development of new mega-cohort studies[7,8] and the creation of large consortia of existing and new cohort and case–control studies.[9] It will be some time before we know how well the suggested megastudies will be able to address the needs for studies of a wide variety of diseases and exposures. For consortial efforts, however, there are emerging signs that this may be an effective paradigm to pursue.

In 2002, investigators for 7 independently funded, case–control studies of non-Hodgkin lymphoma formed the InterLymph Consortium to collaboratively pursue genetic susceptibility and gene–environment interactions for this disease in 3,600 cases and 4,000 controls. Because of the epidemiology of non-Hodgkin lymphoma, the initial focus was on genes related to immune function.[10] Figure 2 illustrates early findings from this effort for 2 genes demonstrating the value of this approach. The TNF gene, and specifically its 308A variant, looks like a very good candidate for further pursuit in both epidemiologic and laboratory studies based on a highly significant pooled estimate resulting from a consistently positive association in each of the 7 studies. Indeed, in the pooled data, compared with the homozygous wild type, the odds ratio for the heterozygote is 1.3, and for the homozygous variant, 1.7. Conversely, data for the IL1A gene indicate that further pursuit may not be warranted, a conclusion that would have been a very long time in coming, if ever, if each study had pursued its own genotyping strategy and published separately over time.

Consortia are also being invoked to meet the challenges presented by the opportunities for whole genome analysis. In 2000, the National Cancer Institute formed the Cohort Consortium, a coalition of 23 large population cohorts studying cancer, which had or were collecting DNA.[11] Nine of the largest cohorts formed a collaboration focused on breast and prostate cancers.[12] Currently, this Consortium is conducting a whole genome analysis of prostate cancer. The multiple studies and large number of cases allow a comprehensive scan of over 500,000 SNPs in one study to be followed by a series of sequential validation/replication efforts in the other cohorts for the emerging candidates (Fig. 3). This strategy will allow a generous definition of good candidates from the scan, to retain good statistical power, but with a robust replication effort to weed out the majority of these that are false-positives to uncover the best candidate genes for further pursuit. The best candidates from this process can then be taken back into the Consortium to look for evidence of a gene–environment interaction.

In summary, "big" science has progressively become an important part of both classic epidemiologic and, more prominently, molecular epidemiologic approaches. It is clear that this trend has not been capricious but has been driven by

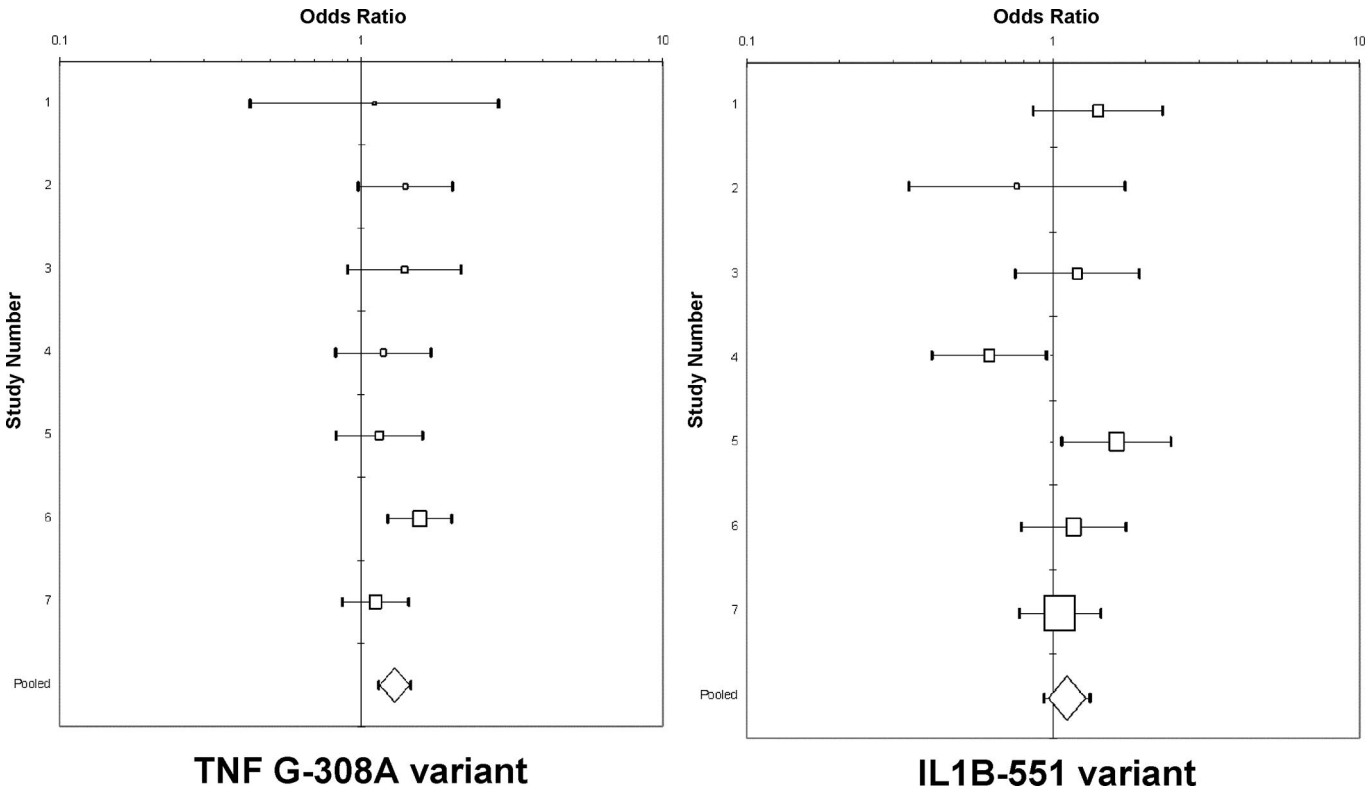

FIGURE 2. Odds ratios for 2 genetic variants and risks of non-Hodgkin lymphoma by individual study and pooled for the InterLymph Consortium.[10]

scientific requirements to meet specific objectives and to take advantage of opportunities resulting from marked scientific and technologic advances. Along with remarkable scientific opportunities afforded by these studies have come significant challenges not only scientific challenges, but to the culture of our discipline as well. Included are issues of opportunities for junior investigators, appropriate credit for scientific contributions in the context of team science, related issues of tenure and promotion, and opportunities for innovation and independence. Others will be addressing these issues in this series of articles. I will offer only 2 brief observations from the perspective of a review of the history of "big" science. First, other disciplines that have incorporated major "big" science components, including physics, astronomy, and genomics, are surviving these challenges to scientific culture and adopting solutions. Given the community orientation of epidemiologists, I am confident that we will even more readily do so. In addition, although a very visible component, "big" science is still only a small part of our discipline. The majority of important epidemiologic questions is, and will continue to be, best addressed by taking innovative and creative advantage of "natural experiments" in the manner practiced since John Snow. Indeed, a similar balance has prevailed in other disciplines as well. Although large collaborations (authorship lists of several hundreds) continue to be prominent in physics, in the past month, the *Journal of Physics B: Atomic, Molecular and Optical Physics* published 58 papers. The average number of authors per paper was 3 with only 4 papers having more than 5 authors.

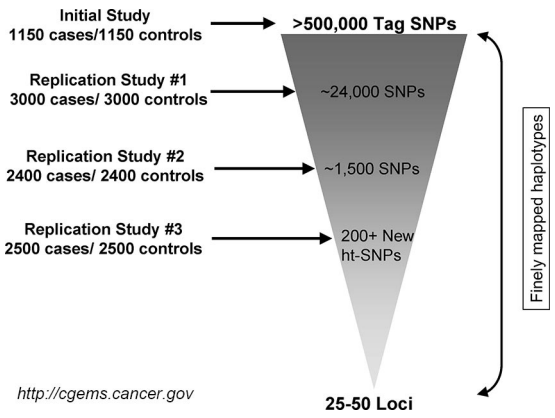

FIGURE 3. Replication strategy for a whole genome analysis study of prostate cancer within the Cohort Consortium.

REFERENCES

1. Howe GR, Burch JD, Miller AB, et al. Artificial sweeteners and human bladder cancer. *Lancet*. 1977;2:578–581.
2. Hoover RN, Strasser PH. Artificial sweeteners and human bladder cancer. Preliminary results. *Lancet*. 1980;1:837–840.
3. Hoover RN. Cancer—nature, nurture, or both. *N Engl J Med*. 2000;343:135–136.

4. Garcia-Closas M, Rothman N, Lubin J. Misclassification in case-control studies of gene–environment interactions: assessment of bias and sample size. *Cancer Epidemiol Biomarkers Prev*. 1999;8:1043–1050.
5. Terry PD, Goodman M. Is the association between cigarette smoking and breast cancer modified by genotype? A review of epidemiologic studies and meta-analysis. *Cancer Epidemiol Biomarkers Prev*. 2006;15:602–611.
6. Wacholder S, Chanock S, Garcia-Closas M, et al. Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J Natl Cancer Inst*. 2004;96:434–442.
7. Collins FS. The case for a US prospective cohort study of genes and environment. *Nature*. 2004;429:475–477.
8. Potter JD. Toward the last cohort. *Cancer Epidemiol Biomarkers Prev*. 2004;13:895–897.
9. National Cancer Institute. Strategic investments in molecular epidemiology. *The Nation's Investment in Cancer Research.* NIH Publication No. 05–5612. December 2004:48–51.
10. Rothman N, Skibola CF, Wang SS, et al. Genetic variation in TNF and IL10 and risk of non-Hodgkin lymphoma: a report from the InterLymph Consortium. *Lancet Oncol*. 2006;7:27–38.
11. Division of Cancer Control and Population Science, Epidemiology and Genetics Research Program. Consortium of Cohorts. 2006. Available at: http://epi.grants.cancer.gov/Consortia/cohort.html. Accessed August 21, 2006.
12. Hunter DJ, Riboli E, Haiman CA, et al. A candidate gene approach to searching for low-penetrance breast and prostate cancer genes. *Nat Rev Cancer*. 2005;5:977–985.

